Privacy and Confidentiality Protection Overview

Simson L. Garfinkel
Senior Scientist, Confidentiality and Data Access
U.S. Census Bureau

May 2, 2019
National Advisory Committee on Racial, Ethnic and Other Populations Spring 2019 Meeting

The views in this presentation are those of the author, and not those of the U.S. Census Bureau.

Acknowledgments

This presentation is based on numerous discussions with

John Abowd (Chief Scientist)

Dan Kifer (Scientific Lead)

Salil Vadhan (Harvard University)

Robert Ashmead, Philip Leclerc, William Sexton, Pavel Zhuravlev (US Census Bureau)

Statistical agencies collect data under a *pledge of confidentiality.*

We pledge:

- Collected data will be used *only for statistical purposes*.
- Collected data will be kept *confidential*.
- Data from individuals or establishments *won’t be identifiable in any publication.*

Fines and prison await any Census Bureau employee who violates this pledge.

Statistical agencies are trusted curators.

Respondents

Confidential Database

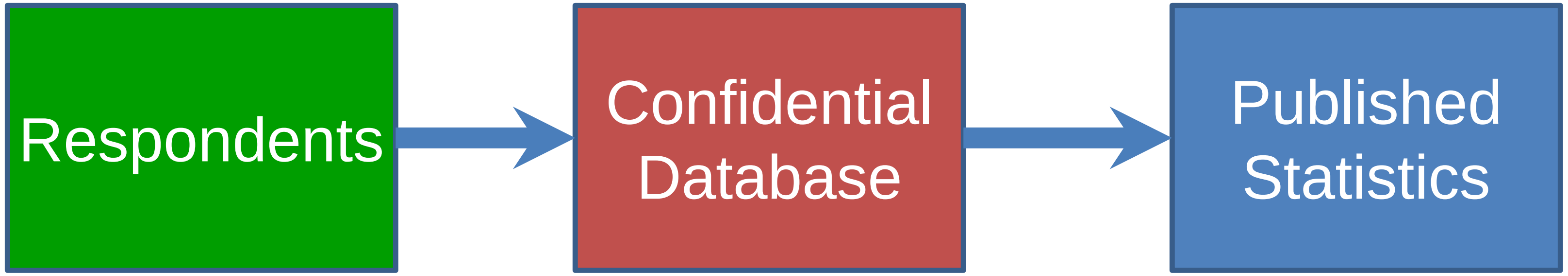
Age Sex Race/MS
8 FBS
18 MWS
24 FWS
30 MWM
36 FBM
66 FBM
84 MBM

Published Statistics

	#	Median Age	Mean Age
Total	7	30	38
Women	4	30	33.5
Male	3	30	44
Black	4	51	48.5
White	3	24	24
Married	4	51	54
Black F	3	36	36.7

We now know “trusted curator” model is more complex.

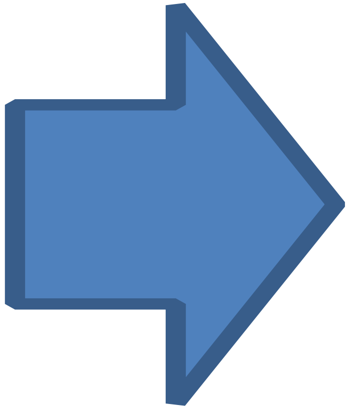
Every data publication results in some privacy loss.



Publishing too many statistics results in the compromise of the entire confidential database.

Consider the statistics from a single household

24 yrs Female White Single (24 FWS)



	Count	Median	Mean
Total	1	24	24
# Female	1	24	24
# white	1	24	24
Single	1	24	24
White F	1	24	24

Publishing statistics for this household alone would result in an improper disclosure.

24 yrs Female White Single (24 FWS)

	Count	Median	Mean
Total	(D)	(D)	(D)
# Female	(D)	(D)	(D)
# white	(D)	(D)	(D)
Single	(D)	(D)	(D)
White F	(D)	(D)	(D)

(D) Means suppressed to prevent an improper disclosure.

In the past, statistical agencies aggregated data from many households together into a single publication.

	Count	Median Age	Mean Age
Total	7	30	38
# Female	4	30	33.5
# male	3	30	44
# black	4	51	48.5
# white	3	24	24
Married	4	51	54
Black F	3	36	36.7

We now know that this publication can be reverse-engineered to reveal the confidential database.

66 FBM & 84 MBM

30 MWM & 36 FBM

8 FBS

18 MWS

24 FWS

	Count	Median	Mean
Total	7	30	38
# Female	4	30	33.5
# male	3	30	44
# black	4	51	48.5
# white	3	24	24
Married	4	51	54
Black F	3	36	36.7

This table can be expressed by 164 equations.
Solving those equations takes
0.2 seconds on a 2013 MacBook Pro.

Faced with “database reconstruction,” statistical agencies have just two choices.

Option #1: Publish fewer statistics.

Option #2: Publish statistics with less accuracy.

The problem with publishing fewer statistics: it’s hard to know how many statistics is “too many.”

Solution #1
8 FBS
18 MWS
24 FWS
30 MWM
36 FBM
66 FBM
84 MBM

Solution #2
2 FBS
12 MWS
24 FWS
30 MBM
36 FWM
72 FBM
90 MBM

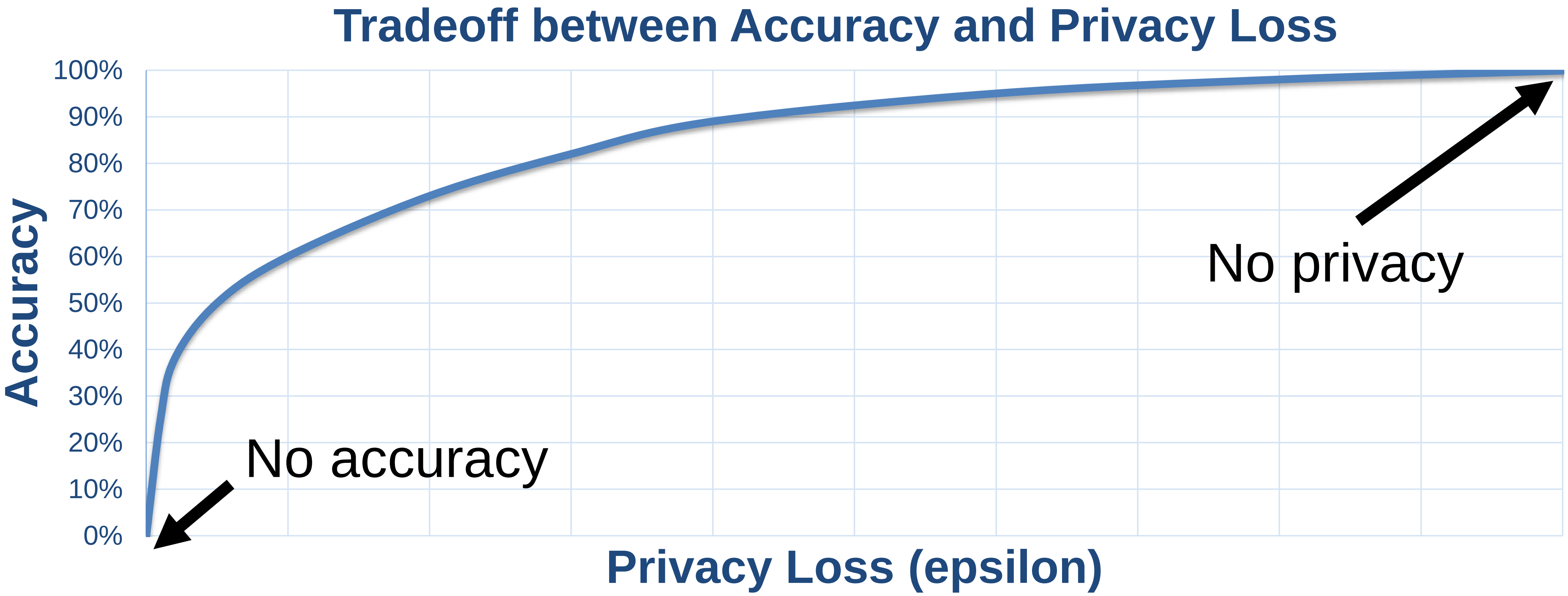
	Count	Median	Mean
Total	7	30	38
# Female	4	30	33.5
# male	3	30	44
# black	4	51	48.5
# white	3	24	24
Single	[redacted]	[redacted]	[redacted]
Married	4	51	54
Black F	[redacted]	[redacted]	[redacted]
Black M	[redacted]	[redacted]	[redacted]
White M	[redacted]	[redacted]	[redacted]
White F	[redacted]	[redacted]	[redacted]

Faced with "database reconstruction," statistical agencies have just ~~two~~ one choice.

~~Option #1: Publish fewer statistics.~~

Option #2: Publish statistics with less accuracy.

Differential privacy gives us a mathematical approach for balancing accuracy and privacy loss.



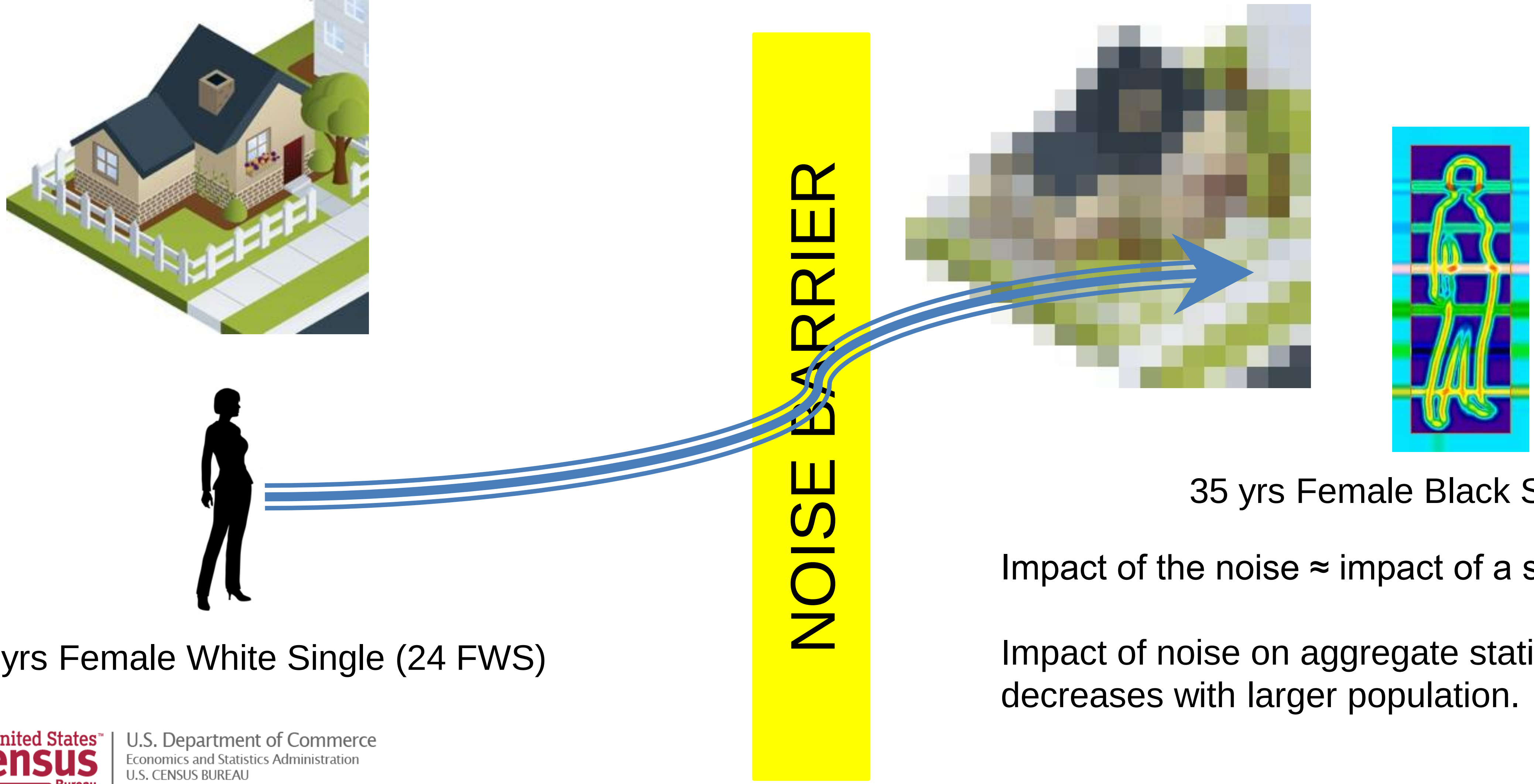
“Differential privacy” is really two things

1 – A mathematical definition of privacy loss.

2 – Specific mechanisms that allow us to:

- ✓ *Add the smallest amount of noise necessary for a given privacy outcome*
- ✓ *Structure the noise to have minimal impact on the more important statistics*

Differential privacy — the big idea:
Use “noise” to create uncertainty about private data.



NOISE BARRIER

24 yrs Female White Single (24 FWS)

35 yrs Female Black Single (35 FBS)

Impact of the noise ≈ impact of a single person

Impact of noise on aggregate statistics decreases with larger population.

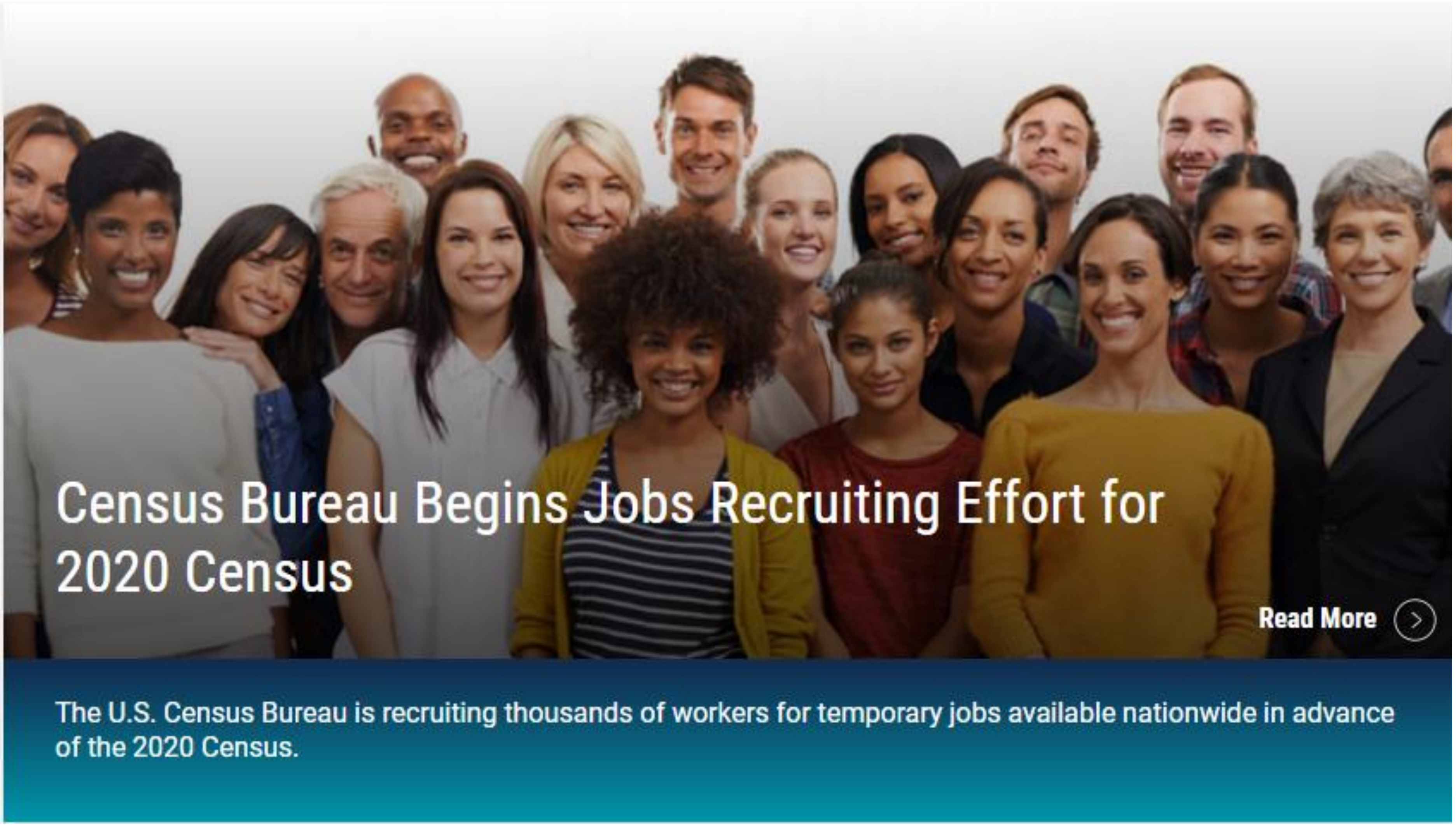
Understanding the impact of “noise:”
(Statistics based on 10,000 experiments, epsilon=1.0)

		50% runs	95% runs
 1 person age 22	NOISE BARRIER	Median(age): 9 → 73	Median(age): 0→ 104
 10 people, all age 22		Median(age): 17 → 61	Median(age): 0→ 103
 100 people, all age 22		Median(age): 21 → 22	Median(age): 21→ 22

The noise also impacts the person counts.

		50% runs	95% runs
 1 person age 22	NOISE BARRIER	Median(age): 9 → 73 # people: -9 → 11	Median(age): 0→ 104 # people: -29 → 30
 10 people, all age 22		Median(age): 17 → 61 # people: 0 → 20	Median(age): 0→ 103 # people: -19 → 38
 100 people, all age 22		Median(age): 21 → 22 # people: 90 → 110	Median(age): 21→ 22 # people: 71 → 129

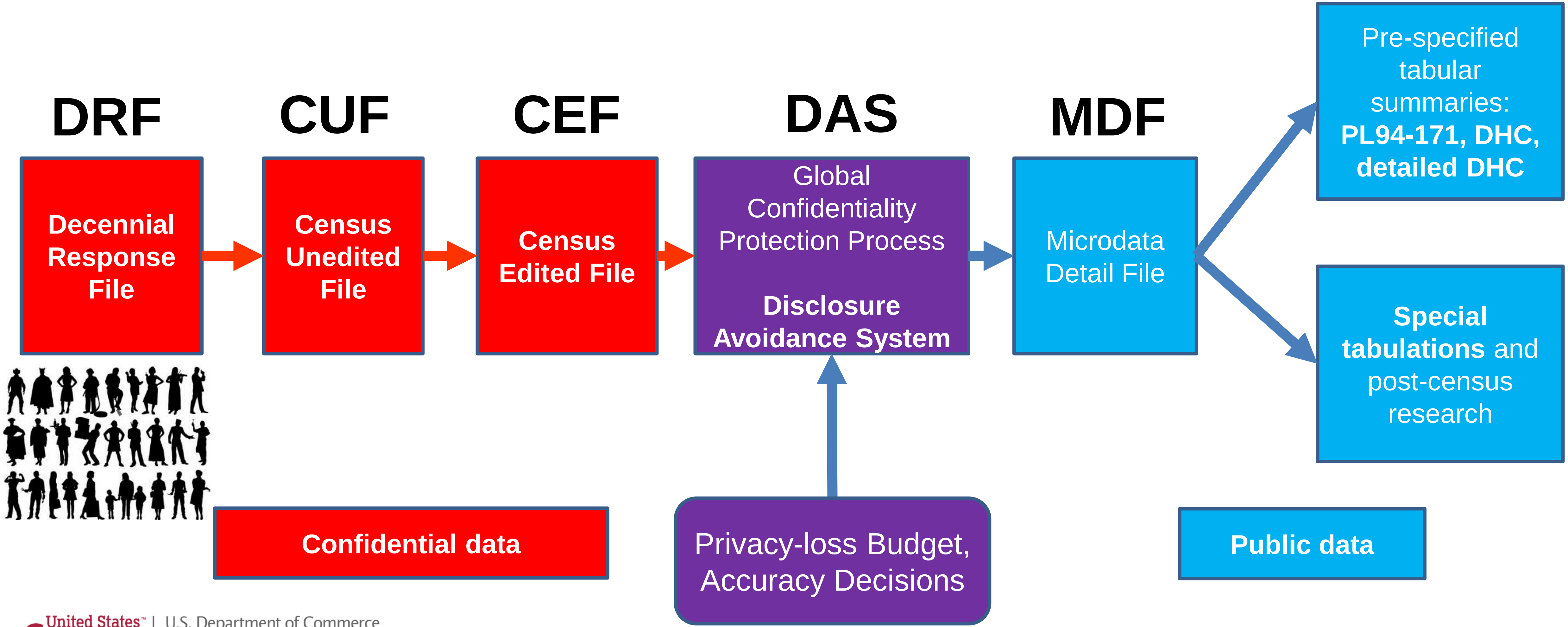
The 2020 census and differential privacy



Census Bureau Begins Jobs Recruiting Effort for 2020 Census

The U.S. Census Bureau is recruiting thousands of workers for temporary jobs available nationwide in advance of the 2020 Census.

The Disclosure Avoidance System allows the Census Bureau to enforce global confidentiality protections.



DRF

Decennial Response File

CUF

Census Unedited File

CEF

Census Edited File

DAS

Global Confidentiality Protection Process

Disclosure Avoidance System

MDF

Microdata Detail File

Pre-specified tabular summaries: **PL94-171, DHC, detailed DHC**

Special tabulations and post-census research

Confidential data

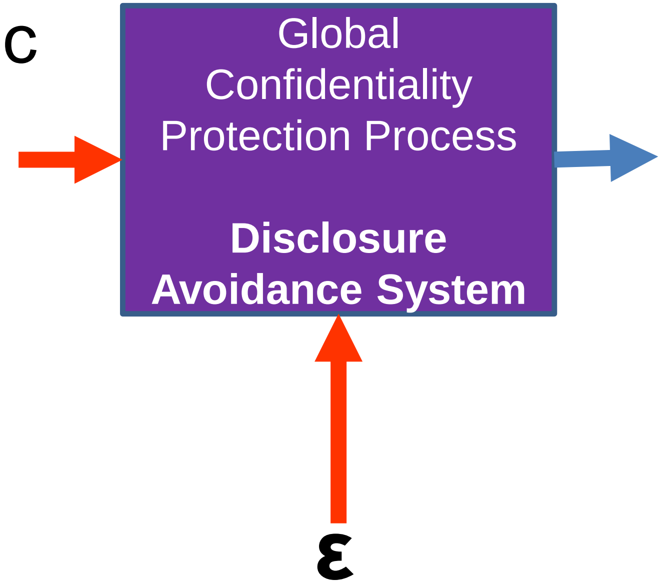
Privacy-loss Budget, Accuracy Decisions

Public data

The Disclosure Avoidance System relies on injects formally private noise.

Advantages of noise injection with formal privacy:

- Transparency: the details can be explained to the public
- Tunable privacy guarantees
- Privacy guarantees do not depend on external data
- Protects against accurate database reconstruction
- Protects every member of the population


Global Confidentiality Protection Process
Disclosure Avoidance System
ε

Challenges:

- Entire country must be processed at once for best accuracy
- Every use of confidential data must be tallied in the *privacy-loss budget*

There was no off-the-shelf system for applying differential privacy to a national census

We had to create a new system that:

- Produced higher-quality statistics at more densely populated geographies
- Produced consistent tables

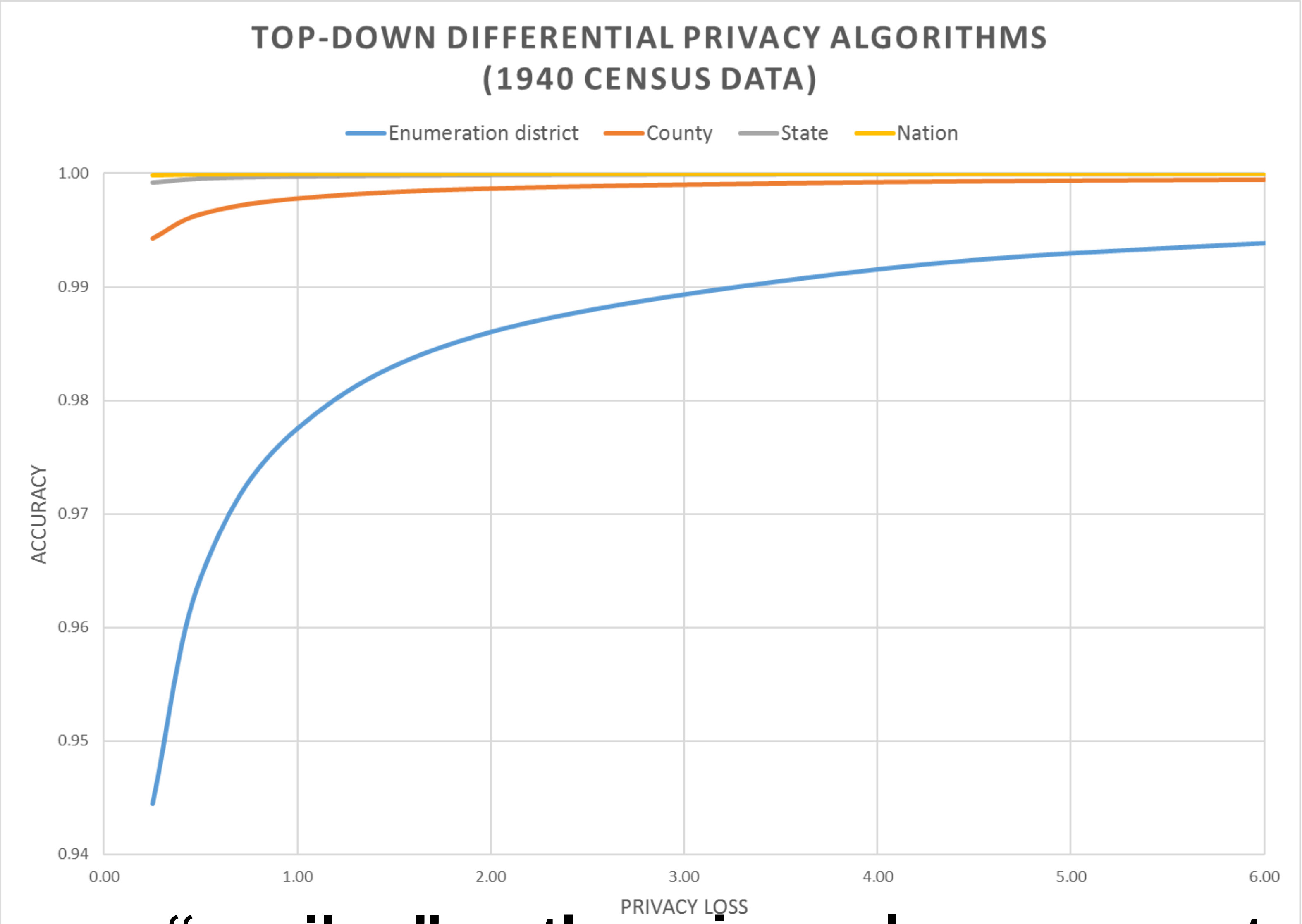
We created new differential privacy algorithms and processing systems that:

- Produce highly accurate statistics for large populations (e.g. states, counties)
- Create privatized microdata that can be used for any tabulation without additional privacy loss
- Fit into the decennial census production system

The 2020 DAS produces highly accurate data when blocks are aggregated into tracts

99% accuracy

95% accuracy



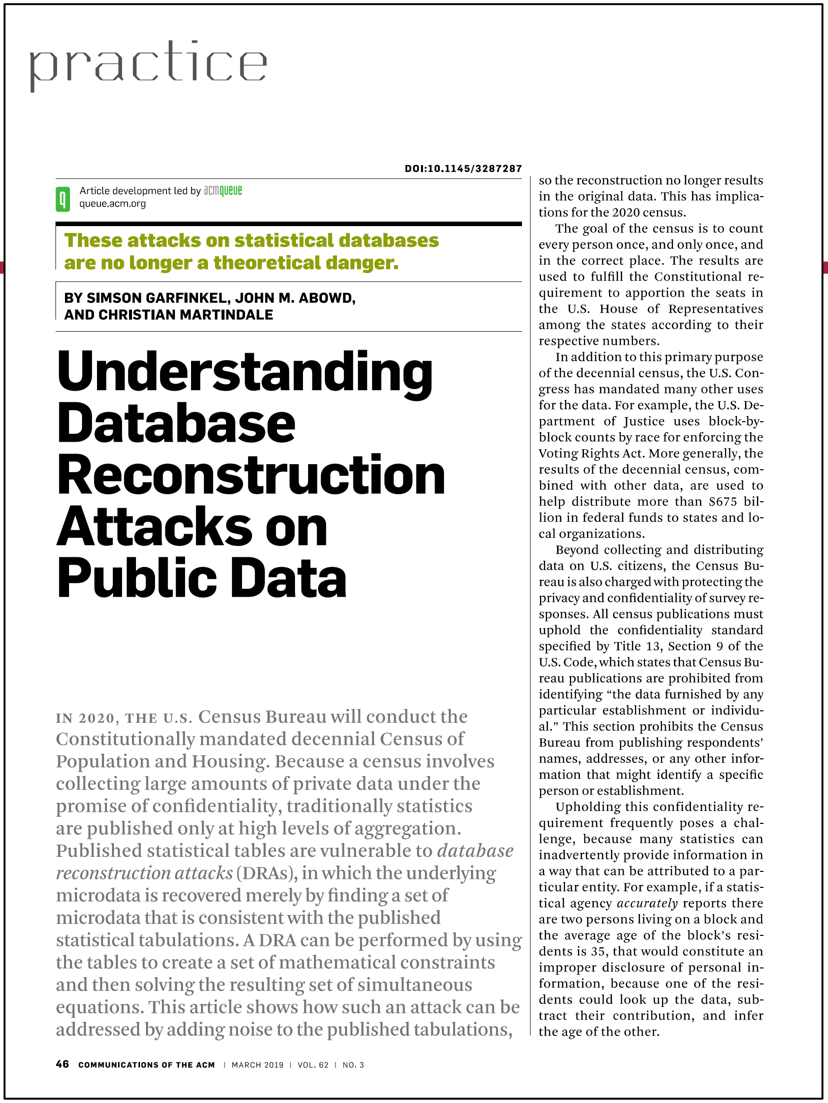
“epsilon” — the privacy loss parameter

Two public policy choices:

What is the correct value of epsilon?

Where should the accuracy be allocated?

For more information…



practice

DOI:10.1145/3287287

Article development led by acmqueue
queue.acm.org

These attacks on statistical databases are no longer a theoretical danger.

BY SIMSON GARFINKEL, JOHN M. ABOWD, AND CHRISTIAN MARTINDALE

Understanding Database Reconstruction Attacks on Public Data

IN 2020, THE U.S. Census Bureau will conduct the Constitutionally mandated decennial Census of Population and Housing. Because a census involves collecting large amounts of private data under the promise of confidentiality, traditionally statistics are published only at high levels of aggregation. Published statistical tables are vulnerable to *database reconstruction attacks* (DRAs), in which the underlying microdata is recovered merely by finding a set of microdata that is consistent with the published statistical tabulations. A DRA can be performed by using the tables to create a set of mathematical constraints and then solving the resulting set of simultaneous equations. This article shows how such an attack can be addressed by adding noise to the published tabulations, so the reconstruction no longer results in the original data. This has implications for the 2020 census.

The goal of the census is to count every person once, and only once, and in the correct place. The results are used to fulfill the Constitutional requirement to apportion the seats in the U.S. House of Representatives among the states according to their respective numbers.

In addition to this primary purpose of the decennial census, the U.S. Congress has mandated many other uses for the data. For example, the U.S. Department of Justice uses block-by-block counts by race for enforcing the Voting Rights Act. More generally, the results of the decennial census, combined with other data, are used to help distribute more than $675 billion in federal funds to states and local organizations.

Beyond collecting and distributing data on U.S. citizens, the Census Bureau is also charged with protecting the privacy and confidentiality of survey responses. All census publications must uphold the confidentiality standard specified by Title 13, Section 9 of the U.S. Code, which states that Census Bureau publications are prohibited from identifying "the data furnished by any particular establishment or individual." This section prohibits the Census Bureau from publishing respondents' names, addresses, or any other information that might identify a specific person or establishment.

Upholding this confidentiality requirement frequently poses a challenge, because many statistics can inadvertently provide information in a way that can be attributed to a particular entity. For example, if a statistical agency *accurately* reports there are two persons living on a block and the average age of the block's residents is 35, that would constitute an improper disclosure of personal information, because one of the residents could look up the data, subtract their contribution, and infer the age of the other.

46 COMMUNICATIONS OF THE ACM | MARCH 2019 | VOL. 62 | NO. 3

Communications of ACM March 2019
Garfinkel & Abowd

Can a set of equations keep U.S. census data private?

By Jeffrey Mervis

Science

Jan. 4, 2019 , 2:50 PM

http://bit.ly/Science2019C1

United States™ Census Bureau | U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

More Background on the 2020 Disclosure Avoidance System

September 14, 2017 CSAC (overall design)
https://www2.census.gov/cac/sac/meetings/2017-09/garfinkel-modernizing-disclosure-avoidance.pdf

August, 2018 KDD’18 (top-down v. block-by-block)
https://digitalcommons.ilr.cornell.edu/ldi/49/

October, 2018 WPES (implementation issues)
https://arxiv.org/abs/1809.02201

October, 2018 *ACMQueue* (understanding database reconstruction)
https://digitalcommons.ilr.cornell.edu/ldi/50/ or
https://queue.acm.org/detail.cfm?id=3295691